

Functional Aspects of Databases: A Research Project Within the Framework of the European Program HCM

Maria Covadonga Fernández-Baizán

Polytechnical University of Madrid

Libia Pérez Jiménez

Technical University of Madrid

Rafael Portaencasa Baeza

Polytechnical University of Madrid

Rosa González Tirados

Polytechnical University of Madrid

Concepción Pérez Liera

University of Oviedo

Eugenio Santos Menéndez

Technical University of Madrid

1. THE HCM PROGRAM

1.1 RATIONALE AND OBJECTIVES

This programme is intended to help increase the human resources available for research and technological development which will be needed by the Member States of the European Community in the coming years. The main aim is the training of research scientists by mobility and the formation of networks.

1.2 TECHNICAL AREAS/ACTIVITIES

The programme will cover all scientific and technological sectors, and areas of the social and human sciences that are able to improve European competitiveness, such as economic and management science, environmental economics, and the interconnections between science, technology and society.

The activities carried out under the programme are:

Activity 1: Fellowships

The development of a system of research fellowships, in particular for young researchers at post-doctoral level,

Activity 2: Networks

The creation and development of scientific and technical co-operation networks across all the regions of the Community, with particular regard to the special needs of less favoured regions.

Activity 3: Large-scale facilities

The organization of series of high level meetings at the cutting edge of scientific or technical knowledge.

2. SHORT DESCRIPTION OF THE PROJECT

2.1 CONTENTS

The motivation for our project is to address the shortcomings of first and second generation DBMSs.

The research topics to be investigated are:

1. Schema Specification and Semantic

We propose to model both the structure and the actual contents of the data-base via graphs. The approach we envisage is to introduce a set of syntactic and semantic constructs for graph-based modelling.

2. Graphical Interfaces

There is a serious lack of theoretical foundations for the new complex interfaces being developed for new generation DBMSs. We propose to investigate the underlying semantics of all aspects of user interaction with the schema and the actual data including querying, updating and browsing.

3. Bulk Types

We propose to develop a general theory of bulk types whose objectives will be to give a precise definition of the concept of a bulk type and to define a common set of operations for querying and updating bulk types including both basic and derived operations.

4. Formal Design Tools

We propose to apply software engineering techniques which emphasise the specification and design phases

of the life-cycle to the area of database specification and design. In particular, we will investigate the transformation of a constraint schema to a transaction schema using the formalism of abstract machines to provide the underlying semantics of transformations.

5. Experiments

Finally, although the proposed research is mainly of a theoretical nature we plan to carry out several experiments in order to test the applicability of our results. These experiments will be carried out in the areas of geographical information systems and software engineering databases.

2.2 BUDGED

The total cost of the Project is 865.458 ECU.

2.3 DURATION

Three years

3. THE HCM-FAD TEAM AT THE UPM

Nine teams pertaining to several european countries are working together in this Project.

Each team has a team-leader and the staff involved included researchers (post graduate people), technicians (undergraduate people) and others (secretary, etc.)

Our item is leaded by Maria C. Fernandez and the staff involved involves six researches, twelve technicians and a secretary.

We are charged of the following research topics:

Schema Specification and Semantics

In recent years the object-oriented approach to databases has received considerable attention and many new database systems have appeared. However, the absence of a single standard object-oriented data model [13] has resulted in much confusion as to what object-oriented database means. A few papers, including [4] and [11], have dispelled some of the confusion, by clarifying the so-called core object-oriented concepts. However, in our opinion, the question of formal semantics for object-oriented databases remains unanswered [13]. Focusing on schema specification and formal semantics issues will be one of the aims of the consortium.

The data entities encountered in advanced database

applications are usually complex, with arbitrary properties and relationships. As a consequence, many researchers have argued that, at the conceptual level, entities should be structured as graphs, in which nodes represent data entities while edges represent relationships (of various kinds) amongst entities. However, in most current systems, the graphical aspect does not go beyond the representation of data, and users of such systems are forced to access in a flat way data which is conceptually represented as a graph. We believe that these graphs should be directly apparent to the users. In other words, both the structure and actual contents of the database should be represented via graphs.

The approach envisaged herein is to introduce a set of syntactic and semantic constructs for graph-based data modelling, using essentially three concepts: the concept of graph, the concept of partial function, and the concept of extension of a graph into sets and partial functions. (By extension of a graph, we mean a function that associates each node of the graph with a set and each edge thereof with a partial function.) The schema is seen as a graph built up from other elementary graphs, called structures, and the semantics of a schema can be viewed as an extension of the graph satisfying a set of path constraints. We note that this way of specifying a schema follows along the lines of algebraic specifications: the schema-graph can be seen as the signature and the schema-constraints can be seen as the axioms of the specification.

Work along these lines has been pursued within the Esprit Basic Research Action FIDE, in which some partners of the present consortium have participated [12]. The results obtained in FIDE will be integrated into the present approach.

Graphical Interfaces

In the area of database systems the facility that allows a user to interact with the database, also known as the user interface, has always been of particular concern. Recently, this concern has greatly increased as a result of the greater complexity of new database applications. Indeed, users of these new applications are specific domain experts rather than application programmers. As advanced database applications become more complex, there is a need for powerful and yet easy to use interfaces for the new DBMSs

developed to solve the above problems. Although a lot has already been published on how such interfaces should look like, there is still a serious lack of theoretical foundations in this area. Such theoretical foundations will be one of the aims of our project.

In the data model that we propose, structure as well as actual contents of the database are represented as graphs. It is therefore natural to want to develop graph-based methodologies and formalisms that allow the user to specify easily common database tasks, and the DBMS to present the results of such tasks in a visually appealing manner. More precisely, the following aspects will be investigated:

Schema specification,
Data definition,
Querying and updating,
Schema modification,
Browsing and viewing.

Our investigations will be guided by state-of-the-art results in the area [8,19], and will have as the driving idea the following widely accepted principle: in a system dealing with complex structures, the user should be able to handle the data elements directly and in a visually appealing manner.

It will be of particular interest to us to investigate the underlying semantics of user interaction with data. This will involve incorporating into the interface: deductive functionalities, the representation of missing and incomplete information and the applicability of non-deterministic semantics (e.g. when choosing a member from a set). In addition, we intend to carry out a preliminary investigation of upgrading the semantics of querying, updating and browsing to a distributed environment.

Bulk Types

Traditionally, applications using a database are programmed in a host programming language. The structure of the persistent data is defined using the abstraction mechanism of the data model, and procedural aspects are then coded in the programming language. As a consequence, new types of complex data such as relations, nested relations, class extents, bags and maps, have been added to the traditional types that one usually finds in a programming language. In the context of advanced database applications, the term bulk type is used to refer to homogeneous collections of complex data.

There is a striking similarity between the operations that one can perform on these bulk types.

Moreover, the following statements are true for all bulk types:

- There is a bulk type with no components,
- For every value there is a bulk type with only that value as a component,
- Bulk types can be combined to create new bulk types.

So the question is: is it possible to treat these different kinds of bulk types uniformly? In other words, is it possible to have a type constructor and polymorphic functions that take bulk types as parameters? Several answers to this question have been proposed. For example, Moggi [16] proposes the categorical concept of monad as a type constructor capable of capturing common operational properties of bulk types used in classical programming languages. Atkinson, Ricard and Trinder [5] propose that the type "map" be used for large scale data. On the other hand, several new programming languages, such as DBPL [15], Galileo [3] and Napier [17] provide features for manipulating bulk types. As a consequence, several authors have recognized the necessity for a formal treatment of bulk types [10].

The consortium aims to develop a general theory of bulk types, the main objectives being:

- A precise definition of the concept of bulk type that encompasses all the known collection types used in the areas of databases and programming languages,
- A common set of basic operations on bulk types and a common query and update notation with uniform semantics over a variety of bulk types,
- A rich set of derived operations for each bulk type, reflecting its standard usage.

3. THE EUROPEAN DIMENSION OF THE PROPOSAL

3.1 ADVANCEMENT OF SCIENCE

The thrust of the Proposal is to lay the theoretical foundations upon which the future generations of Data Base Management Systems (DBMSs) will be built. Currently the market place, dominated by relational DBMSs, is almost exclusively non-European. (The only recent exception regarding the emerging object-oriented DBMSs is O2.) This

situation puts European economies at a disadvantage and in the long term will erode European competitiveness. The FAD proposal can be viewed as the seed corn for the DBMSs that will hopefully appear early next century.

It is hoped that European software Houses, pooling their resources, will avail themselves of the results emanating out of the FAD proposal; they could use these results to develop powerful DBMSs that can easily be used in the new application areas referred to in the introduction of the Proposal. The competition in the development of DBMSs is fierce and only Community-wide partnerships have any chance of success. This will per se induce integration.

Finally, the research to be pursued in the context of the FAD proposal is at the cutting edge of technology and as such will enhance the science base in Europe.

3.2 CONTRIBUTION TO HCM OBJECTIVES

"Databases" are at the heart of information systems which is an important strategic area, since the need to access information will be even more imperative in the future for most application areas, e.g. commercial, manufacturing, etc. Thus by creating an expertise in areas such as Graphical User Interfaces and a solid theoretical basis for their understanding, the wealth creating industries of Europe will gain competitive advantages vis a vis their rivals on the world stage.

In view of the number of Ph.D students and Post-doctoral Research Assistants to be involved in the FAD proposal there will be a qualitative and quantitative increase in the human capital of the Member States. These young researchers will be the scientific leaders who will guide the Community at the part of the next century. It is imperative that this pool of scientifically trained manpower is spread across the Community, a goal this proposal is endeavouring to achieve.

3.3 SOCIAL AND ECONOMIC IMPACT

The large number of researchers involved, especially young scientists, will have to travel to the various sites, in the Member States of the participating teams, meeting colleagues and co-workers from other Community countries; it is anticipated that this will help in breaking down national barriers and thus the European ideal, as perceived by the founding fathers

of the Community, will be fostered among the scientific community. By realising this, enhancement of the cohesion of the Community will ensue and, in particular, that of the academic institutions on a Community-wide basis.

The economic benefits to the less-favoured regions of the Community cannot be quantified, however there will be substantial benefits in the long term, resulting from technological training. In particular, the supply side of their local economy will improve through such technological training. This obviously applies to the developed areas of the Community.

REFERENCES

- [1] Abiteboul S., Bidoit N., "Non First Normal Form Relations: An Algebra Allowing Data Restructuring", *J. Computer and Aystem Sciences*, 33-3, 1986.
- [2] Abiteboul S., Vianu V., "Transaction Languages for Database Update and Specification", Research Report No. 715, INRIA, Rocquencourt, France, 1987.
- [3] Albano A., Cardelli L., Orsini R., "Galileo: a Strongly Typed Conceptual Language", *ACM Trans. on Database Systems*, 10-2, 1985.
- [4] Atkinson M., Bancilhon F., et al., "The Object-Oriented Database System Manifesto", 1st Int. Conf. on Deductive and Object-Oriented Database, Kyoto, Japan, 1989.
- [5] Atkinson M., Ricard P., Trinder P., "Bulk Types for Large-Scale Programming", *Lecture Notes in Computer Science*, 504, 1991.
- [6] Bancilhon F., "Object-Oriented Database Systems", Int. Symposium on Principles of Database Systems, ACM SIGACT-SIGMOD-SIGART, Austin, Texas, 1988.
- [7] Banerjee J., Chou H.T., Garza J.G., Kim W., Woelk D., Ballou N., Kim H.J., "Data Model Issues for Object Oriented Applications", *ACM Trans. on Information Systems*, 3-1, 1987.
- [8] Boursier P., Mainguenaud M., "Spatial Query Languages: Extended SQL vs. Visual Languages vs. Hypermedia", 5th Int. Symposium on Spatial Data Handling, Charleston, Virginia, 1992.
- [9] Hull R., King R., "Semantic Data Modelling: Survey, Applications and Research Issues", *ACM Computing Surveys*, 19-3, 1987.
- [10] Hull R., Su J., "On Bulk Data Types Constructors and Manipulation Primitives: a

work for Analyzing Expressive Power and Complexity", 2nd Int. Workshop on Database Programming Languages, Salishan, Oregon, 1989.

[11] Kim W., "Object-Oriented Databases: Definition and Research Directions", IEEE Trans. on Knowledge and Data Engineering, 2-3, 1990.

[12] Lellahi S.K., Spyrtos N., "Towards a Categorical Data Model supporting Structured Objects and Inheritance", Lecture Notes in Computer Science, 504, 1991.

[13] Maier D., "Why Isn't There an Object-Oriented Data Model?", IFIP 11th World Computer Conf., San Francisco, 1989.

[14] Maier D., Stein J., Otis A., Purdy A., "Development of an Object-Oriented Management System", ACM OOPSLA Int. Conf., Portland, Oregon, 1986.

[15] Matthes F., Schmidt J.W., "Bulk Types: Built-In or Add-On?", 3rd Int. Workshop on Database Programming Languages: Bulk Types and Persistent Data, Nafplion, Greece, 1991.

[16] Moggi E., "Computational Lambda-Calculus and Monads", IEEE Symp. on Logic in Computer Science, Asilomar, California, 1989.

[17] Morrison R., Brown A.L., Connor R.C.H., Dearle A., "The Napier 88 Reference Manual", Research Report 77, Univ. of St Andrews, Dept of Computer Science, 1989.

[18] Naqvi S., Tsur S., "A Logical Language for Data and Knowledge Bases", Computer Science Press, Rockville, Md., 1989.

[19] Paredaens J., P. Peelman, L.Tanca, "G-Log: A Declarative Graphical Query Language", 2nd Int. Conference on Deductive and Object-Oriented Databases, Munich, Germany, 1991.

Maria Covadonga Fernández-Baizán

She was born in Oviedo (Spain) in 1956 and has received the degrees of Electronic Engineering (1978) and Ph. D. in Computer Science (1981), both in the Polytechnical University of Madrid.

Since 1981 she is working in the field of Data Bases. She has published many papers and several books on this topic.

She is, at present, full professor at the Faculty of Informatics of the UPM, where she is head of the

Department of Languages, Systems and Software Engineering.

Since 1983 she has visited, as invited professor, different Universities from several South-american and European countries.

Rosa González Tirados

She was born in Leon (Spain), and received a Ph. D. in Psychology.

As Director of the ICE (Education Science Institute) of the Polytechnical University of Madrid, she is responsible for the introduction and application of new methods and technologies in the educational process; and for evaluating its results.

She has published many papers and several books in the field of Education in Engineering, and has collaborated in many educational projects with South-american countries.

Libia Pérez Jiménez

She was born in Madrid (Spain) in 1951. She received the B.S. degree in Electrical Engineering and the Ph. D. degree in Computer Science in 1977 and 1982, respectively, both from the Technical University of Madrid. Since 1985 she is a Professor of Computer Graphics at the Faculty of Computer Science of the Technical University of Madrid. Her current research interests included advanced topics in Visual Computing.

Concepción Pérez Llera

She was born in Gijón (Spain) in 1959. She received the B.S. degree in Chemistry and the Ph.D. degree in Computer Science in 1981 and 1991, respectively, both from the University of Oviedo. Since 1988 she is a professor of Data Bases and Information Systems at the Faculty of Computer Science of the University of Oviedo. Her current research interest is Semantic Data Models.

Rafael Portaencasa Baeza

He was born in Madrid (Spain) in 1936 and has received the degrees of Electronic Engineering (1961) Computer Scientist (1966) and Ph. D. in Electronic Engineering; all in the Polytechnical University of Madrid.

His scientific work in the field of Computer Science is well known in many American and European countries.

He is, at present, Rector of the Polytechnical University of Madrid.

Eugenio Santos Menéndez

He was born in Madrid (Spain) in 1963. He received the B.S. degree in Computer Science in 1992 from Technical University of Madrid. Since 1992 he is a Professor of Data Bases at the Faculty of Computer Science of the Technical University of Madrid. His current research interests included advanced topics in Object-Oriented Data Bases and Computer Assisted Teaching.